Gépi szövegfeldolgozás nagyméretű párhuzamos korpuszokon

Varga Dániel

Doktori értekezés tézisei

Témavezetők:
Kornai András D.Sc., Lukács András Ph.D.

Eötvös Loránd Tudományegyetem
Informatikai Kar
Információs Rendszerek Tanszék

Informatika Doktori Iskola
Dr. Benczúr András, mat. tud. doktora, egyetemi tanár

Az informatika alapjai és módszertana doktori program
Dr. Demetrovics János, akadémikus

Budapest, 2012

Bevezetés

A tézis új eredményeket mutat be a morfológiai egyértelműsítés, a tulajdonnév-felismerés, a főnévi csoportok kijelölése, a mondat-párhuzamosítás, és a bitext keresés témaköreiben. Az utolsó témakör új, „Web 2.0" fejlemény, de a többi a gépi szövegfeldolgozás (Natural Language Processing, NLP) klasszikus feladata, melyekről az előzmények megtalálhatók a témakör tankönyveiben és kézikönyveiben. A morfológiáról ld. (Mitkov 2005) 2. fejezetét, a tulajdonnév-felismerésről (Jurafsky & Martin 2000) 22.1, a főnévi csoportok kijelöléséről (Manning & Schütze 1999) 10.6.2 és 12.1.2 fejezeteit, a mondat-párhuzamosításról pedig (Jurafsky & Martin 2000) 25.6 fejezetét. A nyelvi erőforrások (Language Resource, LR) az NLP alapvető adathalmazai, legyenek akár nyers (szűretlen) állapotban, mint a legtöbb kiinduló korpusz, akár feldolgozott állapotban, mint a számítógépes szótárak. A disszertációban bemutatott eredmények három csoportba sorolhatók:

(1) Nyelvtechnológiai szoftvereszközök

(2) Nyelvi erőforrások (korpuszok), amelyeket a fenti eszközök felhasználásával előállítottunk magyar és más nyelvekre

(3) Kutatási eredmények, amelyek ezen eszközök és nyelvi erőforrások kiaknázására épülnek

Az 1. fejezet a disszertáció témáját mutatja be, a 2. és 3. fejezet pedig a szükséges matematikai illetve nyelvészeti hátteret. Ezen fejezetek nem tartalmaznak új eredményeket.

A könnyebb követhetőség érdekében a tézisek számozása követi a disszertáció fejezetszámozását (tehát 4-gyel kezdődik). A tézisfüzetben összefoglaljuk a disszertáció főbb eredményeit fejezetről fejezetre haladva, minden fejezetben a fenti hármas tagolást alkalmazva. A bemutatott munka természetéből adódóan erősen kollaboratív, ezért (4) pont alatt minden fejezetnél külön tisztázzuk, hogy az egyes szoftvereszközök és nyelvi erőforrások létrejöttében mi volt a szerző saját munkája. A disszertációban bemutatott minden eszköz és nyelvi erőforrás nyílt forráskódú, és hozzáférhető az alábbi linkeken:

`webcorpus pipeline`	`https://github.com/zseder/webcorpus`
`hunner`, `hunchunk`	`https://github.com/recski/HunTag.git`
`hunalign`	`http://mokk.bme.hu/resources/hunalign`
`Hunglish bitext query`	`http://code.google.com/p/hunglish-webapp`
Szószablya Webkorpusz	`http://mokk.bme.hu/resources/webcorpus`
Gyakorisági szótárak	`http://hlt.sztaki.hu/resources/webcorpora.html`
Hunglish Korpusz	`http://mokk.bme.hu/resources/hunglishcorpus`
JRC-Acquis Corpus	`http://langtech.jrc.it/JRC-Acquis.html`
Szószablya szolgáltatás	`http://szotar.mokk.bme.hu/szoszablya`
Hunglish szolgáltatás	`http://hunglish.hu`

4. Morfológiai egyértelműsítés

Egy morfológiailag komplex nyelven, amilyen a magyar is, a morfológiai elemzés gyakran több lehetséges elemzést is visszaad egy adott szóra. (Például *ment* jelenthet múlt- vagy jelen idejű igét.) A *morfológiai egyértelműsítés* feladata abban áll, hogy kontextus alapján eldöntsük, hogy a szövegszó lehetséges elemzései közül melyik a helyes.

(4.1) Tézis Megépítettünk és kiértékeltünk egy a magyar nyelvű morfológiai egyértelműsítés feladatát megoldó rendszert, state-of-the-art pontosságot elérve.

Eredményeink megmutatják, hogy statisztikai gépi tanuláson alapuló módszerek eredményesen kombinálhatók szabály-alapú morfológiai elemzéssel. Magyarra ezt elsőként (Oravecz & Dienes 2002) demonstrálta, a mi rendszerünk előnye a szótáron kívüli szavak robusztusabb kezelésében rejlik.

(4.2) Tézis Automatikus eszközökkel morfológiailag egyértelműsített webkorpuszt és gyakorisági szótárat építettünk magyar nyelvre. A korpusz 18 millió weboldal alapján épült, és minőségileg legerősebben szűrt változata 589 millió szóból áll 1.22 millió weboldalról, amely a legnagyobb magyar nyelvű nyilvánosan elérhető korpusszá teszi. Összehasonlításképpen a manuálisan összeállított Magyar Nemzeti Szövegtár (Váradi 2002) 188 millió szót tartalmaz , a manuálisan annotált Szeged Korpusz (Csendes et al. 2004) 1.2 millió szót.

(4.3) Morfológiailag egyértelműsített gyakorisági szótárunk fontos segédeszköz magyar nyelvészek (Magyar & Szentgyörgyi 2011, Rácz & Szeredi 2009, Szeredi 2009) és pszicholingvisták (Racsmány et al. 2012, Lukács et al. 2007) számára. A disszertáció 4.9.2 pontjában bemutatjuk a gyakorisági szótár egy pszicholingvisztikai kutatásban való felhasználását (Pléh et al. 2011).

Kutatócsoportunk jelenleg a korpusz és szótár frissített, megnövelt kiadásán dolgozik. Amint azt (Halácsy et al. 2008) és (Zséder et al. 2012) dokumentálja, adatfeldolgozási rendszerünket újraírtuk ehhez a feladathoz, megnövelt feldolgozási sebességgel és az eddiginél is gondosabb adatszűréssel. Rendszerünket felhasználva webkorpuszokat és (morfológiailag egyelőre egyértelműsítetlen) gyakorisági szótárakat építettünk 15 európai nyelvhez.

(4.4) A fejezet legfőbb eredményeit a (Halácsy et al. 2005) és (Kornai et al. 2006) cikkekben közöltük. A disszertáció szerzője egyenrangú közreműködő volt Halácsy Péterrel és Kornai Andrással a (Halácsy et al. 2005) alatt ismertetett rendszerek megalkotásában és kiértékelésében. A webkorpusz és szótár megépítése egy nagyobb kollaboráció keretében történt (Kornai et al. 2006). A szerző számos szövegfeldolgozó és adattisztító komponens létrehozásával járult hozzá ehhez, illetve folyamatban lévő frissítéséhez (Zséder et al. 2012). A közölt pszicholingvisztikai eredményekhez a szerző az entrópia-modellek megalkotásával, illetve a kísérleti adatok statisztikai elemzésével járult hozzá (Pléh et al. 2011).

5. Tulajdonnév-felismerés

A világról való tudásunk nagy része kötődik helyekhez, személyekhez és szervezetekhez, ezért egy szöveg *tulajdonneveinek* azonosítása és kategorizálása kulcsfeladat a természetes nyelvi szöveg feldolgozásakor. Ezt a feladatot angol nevének (Named Entity Recognition)

rövidítésével NER néven ismerjük. (A NER problémakörön belülre tartozik az olyan *metonímikus* olvasatok azonosítása és osztályzása, mint amilyen a 'hely mint esemény' *'Vietnam nagy nemzeti trauma volt.'*, ld. a disszertáció 5.5 fejezetét.)

(5.1) Tézis Megépítettük a `hunner` tulajdonnév-felismerő rendszert magyar nyelvre. A rendszer maximum entrópia modellezést alkalmaz nagyméretű jegyhalmazokon, és standard kiértékelési feladaton sikeres eredményt mutat fel, 95% feletti F-mértékkel (Varga & Simon 2007).

(5.2) Rendszerünk felhasználásával (Solymosi 2007) megalkotta a híranyagokat tartalmazó, automatikusan tulajdonnév-egyértelműsített 73.8 millió szavas Origo NER Korpuszt.

(5.3) A disszertáció 5.6 fejezetében bemutatjuk az Origo NER korpusz egy hálózatkutatási jellegű felhasználását.

(5.4) A `hunner` rendszer közös munka Simon Eszterrel, amelyet a (Varga & Simon 2006) és (Varga & Simon 2007) cikkekben dokumentáltunk. A gépi tanulási architektúra, az implementáció és a kiértékelési keretrendszer a jelen szerző munkái, a tanítási jegyekkel kapcsolatos mérnöki munka közös. A szoftvert később a szerző újraimplementálta Recski Gáborral közösen (Recski & Varga 2009), ennek eredménye a `huntag` rendszer, amely a megfelelő erőforrás birtokában tulajdonnév-felismerésen kívül más feladatok elvégzésére is képes. A szerzőnek nem volt érdemi szerepe az (5.2) és (5.3) pontban említett alkalmazások létrehozásában, ezeket csupán a teljesség kedvéért említjük meg.

6. Főnévi csoportok azonosítása

A gépi szövegfeldolgozás jelenlegi fejlettségi szintjén a mondatok szintaktikai elemzése még nem oldható meg kielégítő pontossággal. Szerencsére számos információkinyerési alkalmazáshoz elegendő a *főnévi csoportok* (noun phrase, NP) azonosítása (NP chunking). A főnévi csoportok azonosítása fontos lépés a teljes mondattani elemzés felé is, egyrészt mivel az NP-k belső szerkezete olyan elterjedt eszközökkel is feltérképezhető, mint a környezetfüggetlen nyelvtanok, másrészt mivel az NP-k megtalálása hozzásegíthet a magasabb szintű mondatszerkezet feltárásához, így például az igék valenciastruktúrájának kitöltésében. Az első magyar NP-azonosító rendszert (Váradi 2003) publikálta, a rendszer 58.78%-os F-mértéket ért el.

(6.1) Tézis Kifejlesztettük a `hunchunk` eszközt, amely versenyképes eredményt ér el a magyar NP-azonosítás feladatán.

(Miháltz 2011) a `hunchunk` rendszert két további NP-azonosítóval együtt értékelte ki: egy szabályalapú magyar NP-azonosítóval (Váradi & Gábor 2004) és a MetaMorpho gépi fordító rendszer által használt mondattani elemzővel (Prószéky et al. 2004). Mint az 1. táblázatból kiderül, a `hunchunk` mindkét szabályalapú rendszernél magasabb pontszámokat ér el. Megjegyezzük, hogy a kiértékelés módszertana a mi rendszerünk irányában erősen elfogult, amennyiben a tesztadat az általunk tanításra használt Szeged Korpuszból került ki (bár tanítóadatunkkal nem is volt átfedésben).
(Hócza 2004) egy szabálytanuláson alapuló NP-azonosító rendszert ismertet, mely 83%-os F-mértéket ér el egy a miénkhez hasonló korpuszon, a kiértékelést tízszeres keresztva-

	Pontosság	Fedés	F-mérték
`hunchunk`	78.67	84.99	81.71
`MetaMorpho`	54.39	61.52	57.73
`NooJ`	37.57	59.28	45.99

1. táblázat. Rendszerünk összehasonlítása két szabályalapú NP-azonosító rendszerrel. (Miháltz 2011)

lidációval végezve. Noha korpuszaink nem teljesen összehasonlíthatók, a keresztvalidációt mi is elvégeztük és 89.30%-os F-mértéket értünk el (pontosság 89.75%, fedés 88.86%).

(6.2) (Recski et al. 2010) magyar és angol főnévi csoportok párhuzamosítását végző rendszert mutat be. A rendszer az NP-k azonosításában a `hunchunk`-ra támaszkodik.

(6.4) Jelen fejezet eredményei Recski Gáborral közösek, melyeket eredetileg (Recski & Varga 2009) és (Recski & Varga 2012) ismertet. A szerző nem vett részt a (Recski et al. 2010)-ben ismertetett munkában.

7. Mondat-párhuzamosítás

A statisztikai alapú gépi fordításnak, valamint a kétnyelvű szótárak automatikus építésének bemenetéül párhuzamos szövegek (bitextek) szolgálnak, melyek ugyanazon anyagokat két nyelven tartalmazzák. A *mondat-párhuzamosítás*, azaz a különböző nyelvű mondatok közti megfeleltetések feltárása egy olyan feladat, mely kulcsfontosságú a gépi fordítás és rokon feladatok által használt adatok előkészítésében. A mondat-párhuzamosítás feladatát végző elterjedt eszközök közé tartozik a `BSA` (Moore 2002) és a `GMA`(Melamed 1998).

(7.1) tézis Kifejlesztettük a `hunalign` rendszert és a hozzá tartozó `partialAlign` előfeldolgozót. Ezeket az különbözteti meg más elterjedt mondat-párhuzamosító rendszerektől, hogy közel egy nagyságrenddel alacsonyabb a futási idejük, hasonló pontosság mellett.

Egy teljes IBM fordítási modell helyett a `hunalign` egy egyszerűbb, de hatékony szótárépítési módszert használ. Ez a módszer két előnnyel bír. Egyrészt a fordítási hasonlóság pontszám rendkívül gyorsan számolható. A `hunalign` szignifikánsan gyorsabb más modern párhuzamosítóknál, és ez a sebességnövekedés többféle módon is kihasználható több tízezer dokumentumból álló párhuzamos korpuszok készítésekor. Másrészt a `hunalign` képes felhasználni egy kétnyelvű szótárat, amennyiben az rendelkezésre áll. A `partialAlign` előfeldolgozó egy gyors és pontos darabolási algoritmust alkalmaz, melynek segítségével elkerülhető, hogy a `hunalign`-t közvetlenül kelljen futtatni 20000 mondatnál hosszabb dokumentumokon, ahol a memóriaigénye szűk keresztmetszetté válna.

Eredményeink eredeti publikálását követően (Krynicki 2006) összehasonlította a `BSA`, `GMA` és `hunalign` rendszereket lengyel-angol szövegen, különböző korpuszokon illetve különböző lemmatizációs beállítások mellett. Úgy találta, hogy a `BSA` gyakran megelőzte a másik két eszközt pontosságban, a `hunalign` azonban minden beállítás mellett első helyen végzett az F-mérték szempontjából, köszönhetően a magasabb fedésnek.

(Abdul Rauf et al. 2012) is mondat-párhuzamosító eszközök teljesítményét hasonlítja össze olyan módon, hogy kimenetükön statisztikai alapú fordítórendszereket (SMT) tanít, melyeket aztán három népszerű metrika szerint kiértékel. Mindhárom metrika alapján elmondható, hogy a `hunalign` „statisztikai holtversenyben" az első helyen végzett a francia-angol fordítási feladaton, amennyiben a legmagasabb pontszámtól legfeljebb a szórás mértékével maradt el.

Mint valamennyi nyelvfeldolgozó eszközünk, a `hunalign` is szabad szoftver. A segítségével készült korpuszok szabadon felhasználhatók tetszőleges célra, beleértve a kereskedelmi célú felhasználást is. Ez megkülönbözteti számos alternatívájától, mely megtiltja az effajta felhasználást.

(7.2) tézis A JRC-Acquis korpusz (Steinberger et al. 2006) egy nagyméretű, soknyelvű párhuzamos korpusz, melyet nyelvtechnológusok nemzetközi csapata épített. 230 idézettel (Google Scholar) a JRC-Acquis az egyik legfontosabb erőforrás a többnyelvű természetes nyelvfeldolgozásban. Elsődleges felhasználási területe statisztikus gépi fordítórendszerek tanítása (pl. (Turchi et al. 2009), de számos egyéb felhasználását is dokumentálták, többek között többnyelvű információ-visszakeresésben (Talvensaari 2008), többnyelvű véleményelemzésben (Bautin et al. 2008) és többnyelvű plágiumfelismerésben (Potthast et al. 2011).

Pontosságának, gyorsaságának, és megengedő licenszének köszönhetően a `hunalign` széles körben elterjedt eszköze a nemzetközi nyelvtechnológus-közösségnek. (Az eszközt bemutató publikáció jelenleg 98 idézetet tudhat magáénak a Google Scholar szerint.) Az általunk ismert öt legnagyobb nyilvános soknyelvű párhuzamos korpusz közül négy a `hunalign` segítségével készült: (i) a szerző közvetlenül részt vett a korábban már említett, meghatározó szerepű JRC-Acquis korpusz (Steinberger et al. 2006) készítésében; (ii) OPUS (Tiedemann 2009); (iii) Parasol (Waldenfels 2011); és (iv) InterCorp (Rosen & Vavřín 2012). Kivételt képez az (v) EUROPARL (Koehn 2005) korpusz, melyet egy korábbi eszközzel párhuzamosítottak, de amelynek szerzője később maga is a `hunalign` használatát javasolja monográfiájában (Koehn 2010).

A legnagyobb angol-magyar párhuzamos korpusz, a Hunglish Korpusz (115 millió szó 4.15 millió mondatpárban, ld. (Varga et al. 2005)), ugyancsak a `hunalign` segítségével készült. Magyarországon ez a korpusz tanítóadatként szolgált a gépi fordításban (Hócza & Kocsor 2006), és a jelentés-egyértelműsítésben (Miháltz & Pohl 2006). Nemzetközi kutatásokban is használták tanítóadatként, a '2008 ACL Workshop on Statistical Machine Translation' (Callison-Burch et al. 2008) és a '2009 EACL Workshop on Statistical Machine Translation' `http://www.statmt.org/wmt09/translation-task.html` versenyeken.

(7.3) A `hunalign` számos szoftverrendszer komponensét képezi. Ezek közül leginkább az `UPlug` és az `LF Aligner` érdemel említést. Az `UPlug` (Tiedemann 2002) egy párhuzamos szövegek feldolgozását szolgáló keretrendszer, ennek segítségével készült az OPUS párhuzamos korpusz is (Tiedemann 2009). Az `LF Aligner` eszköz fő komponenseként a `hunalign` és `partialAlign` eszközöket hivatásos fordítók ezrei használják világszerte fordítómemóriák létrehozására.

(7.4) A `hunalign` és `partialAlign` eszközök tervezése, implementációja és kiértékelése a szerző munkája. A korpuszok létrehozása együttműködésben történt. A Hunglish korpusz esetében (Varga et al. 2005) a szerző végezte a munka meghatározó részét, az adattisztítástól

a csomagolásig. A JRC-Acquis korpusz esetében (Steinberger et al. 2006) a szerző szerepe a mondat-párhuzamosítás munkafázisának elvégzése volt, amelyhez adaptálnia kellett a `hunalign` technológiát.

8. Bitext lekérdezés

A kereshető bitext (párhuzamosított kétnyelvű szöveg) fontos eszköz az emberi fordító számára, aki a segítségével szövegkörnyezetben találhatja meg egyes szavak, szókapcsolatok fordítását. A felhasználói visszajelzés és dokumentumfeltöltés lehetővé tételével a rendszer párhuzamos nyelvi adatok félautomatikus gyűjtését és kézi validációját megvalósító „crowdsource" eszközzé válik.

(8.1) tézis Kifejlesztettük a Hunglish bitext-kereső rendszert. A rendszer két részből, egy offline (aszinkron) és egy online (szinkron) komponensből áll. Az offline komponens bemenete két dokumentum, melyekből mondatpárok halmazát nyeri ki, majd ezeket hozzáadja egy indexhez. Az online komponens felhasználói lekérdezéseket szolgál ki a mondatpárok ezen indexe alapján. A komponensek egy webes alkalmazás részeiként működnek, mely a lekérdezések megválaszolásával akár egyidőben is képes a felhasználók által feltöltött dokumentumokból kinyert mondatpárokkal bővíteni az adatbázist. A felhasználónak lehetősége nyílik szavazni az egyes mondatpárok helyességéről, ezzel javítva a párhuzamos korpusz minőségét. A negatív szavazatokat kapott mondatok azonnal hátrébb kerülnek a találati listákon. A szolgáltatás jelenleg a magyar-angol nyelvpárra áll rendelkezésre, de a szoftver teljesen nyelvfüggetlen.

(8.2) tézis A szolgáltatás, mely egyelőre béta-fázisban van, és nem rendelkezik kiforrott felhasználói felülettel sem, már számottevő mértékű webes jelenléttel rendelkezik. Havonta 10000 felhasználó 150000 lekérdezését válaszolja meg. A látogatók eddig 2000 mondatpárt jelöltek hibásnak, és 37 új dokumentumpárt töltöttek fel, melyekből a korpusz 17000 új mondatpárral bővült.

(8.4) A Hunglish rendszer offline komponense Zséder Attilával közös munka eredménye (Recski et al. 2009). A szoftvert a szerző tervezte. A `hun*` eszközök és más nyelvfeldolgozó kellékek ezen keretrendszerbe való beillesztése szintén a szerző munkájának eredménye. Az online komponens Barna Péter Gergővel együttműködésben készült, eddig publikálatlan. A rendszer architektúrája közös munka eredménye, az implementáció nagyrészt Barna Péter Gergő munkája. A rendszer teljes újraimplementálása korábbi bitext-lekérdező rendszerünknek, mely Halácsy Péter munkája (Halácsy et al. 2004). A reimplementáció során beépített legfontosabb új funkciók az új dokumentumpárok hozzáadásának lehetősége, a duplikátumszűrés, valamint a felhasználói visszajelzés lehetősége.

Angol nyelvű publikációk

Farkas, R., Simon, E., Szarvas, G. & Varga, D. (2007), GYDER: Maxent metonymy resolution, *in* 'Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)', Association for Computational Linguistics, Prague, Czech Republic, pp. 161–164.

Halácsy, P., Kornai, A., Németh, P. & Varga, D. (2008), Parallel creation of gigaword corpora for medium density languages - an interim report, *in* 'Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)', European Language Resources Association (ELRA), Marrakech, Morocco.

Kornai, A., Halácsy, P., Nagy, V., Oravecz, C., Trón, V. & Varga, D. (2006), Web-based frequency dictionaries for medium density languages, *in* 'Proceedings of the EACL 2006 Workshop on Web as a Corpus'.

Pléh, C., Németh, K., Fazekas, J. & Varga, D. (2011), Entropy measures and predictive recognition as mirrored in gating and lexical decision over multimorphemic Hungarian noun forms, *in* 'QMMMD Workshop, University of California, San Diego (Jan. 15-16)'.

Recski, G. & Varga, D. (2009), 'A Hungarian NP-chunker', *The Odd Yearbook* .

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufiş, D. & Varga, D. (2006), The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages, *in* 'Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006)', Genoa, Italy.

Trón, V., Gyepesi, G., Halácsy, P., Kornai, A., Németh, L. & Varga, D. (2005), Hunmorph: open source word analysis, *in* 'Proceedings of the ACL 2005 Workshop on Software'.

Varga, D., Németh, L., Halácsy, P., Kornai, A., Trón, V. & Nagy, V. (2005), Parallel corpora for medium density languages, *in* 'Proceedings of the Recent Advances in Natural Language Processing 2005 Conference', Borovets. Bulgaria, pp. 590–596.

Varga, D. & Simon, E. (2007), 'Hungarian named entity recognition with a maximum entropy approach', *Acta Cybern.* **18**(2), 293–301.

Zséder, A., Recski, G., Varga, D. & Kornai, A. (2012), Rapid creation of large-scale corpora and frequency dictionaries, *in* 'Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)', European Language Resources Association (ELRA), Istanbul, Turkey.

Magyar nyelvű publikációk

Farkas, R., Szeredi, D., Varga, D. & Vincze, V. (2010), MSD-KR harmonizáció a Szeged Treebank 2.5-ben, *in* 'VII. Magyar Számítógépes Nyelvészeti Konferencia', pp. 349–353.

Halácsy, P., Kornai, A., Németh, L., Rung, A., Szakadát, I., Trón, V. & Varga, D. (2004), Hunglish: nyílt statisztikai magyar–angol gépi nyersfordító, *in* 'II. Magyar Számítógépes Nyelvészeti Konferencia', Szegedi Tudományegyetem, pp. 81–84.

Halácsy, P., Kornai, A. & Varga, D. (2005), Morfológiai egyértelműsítés maximum entrópia módszerrel (morphological disambiguation with the maxent method), *in* 'Proc. 3rd Hungarian Computational Linguistics Conf.', Szegedi Tudományegyetem.

Rebrus, P., Kornai, A. & Varga, D. (2012), 'Egy általános célú morfológiai annotáció', *Általános Nyelvészeti Tanulmányok* . to appear.

Recski, G. & Varga, D. (2012), 'Magyar főnévi csoportok azonosítása', *Általános Nyelvészeti Tanulmányok* . to appear.

Recski, G., Varga, D., Zséder, A. & Kornai, A. (2009), Főnévi csoportok azonosítása magyar-angol párhuzamos korpuszban, *in* 'VI. Magyar Számítógépes Nyelvészeti Konferencia', Szegedi Tudományegyetem.

Varga, D. & Simon, E. (2006), Magyar nyelvű tulajdonnév-felismerés maximum entrópia módszerrel, *in* Z. Alexin & D. Csendes, eds, 'IV. Magyar Számítógépes Nyelvészeti Konferencia', Szegedi Tudományegyetem, Szeged, pp. 32–38.

Hivatkozások

Abdul Rauf, S., Fishel, M., Lambert, P., Noubours, S. & Sennrich, R. (2012), Extrinsic evaluation of sentence alignment systems, *in* LREC Workshop on Creating Cross-language Resources for Disconnected Languages and Styles (CREDISLAS)', Istanbul (Turkey).

Bautin, M., Vijayarenu, L. & Skiena, S. (2008), International sentiment analysis for news and blogs, *in* 'Proceedings of the International Conference on Weblogs and Social Media (ICWSM)'.

Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C. & Schroeder, J. (2008), Further meta-evaluation of machine translation, *in* 'Proceedings of the Third Workshop on Statistical Machine Translation', StatMT '08, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 70–106.

Csendes, D., Csirik, J. & Gyimóthy, T. (2004), The Szeged Corpus: A POS tagged and syntactically annotated Hungarian natural language corpus, *in* Text, Speech and Dialogue: 7th International Conference, TSD', pp. 41–47.

Hócza, A. (2004), Noun phrase recognition with tree patterns', *Acta Cybern.* **16**(4), 611–623.

Hócza, A. & Kocsor, A. (2006), Hungarian-English machine translation using genpar, *in* 'Proceedings of the 9th international conference on Text, Speech and Dialogue', TSD'06, Springer-Verlag, Berlin, Heidelberg, pp. 87–94.

Jurafsky, Daniel & Martin, James, (2000), *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*, Prentice Hall; 1st edition.

Koehn, P. (2005), Europarl: A Parallel Corpus for Statistical Machine Translation, *in* 'Conference Proceedings: the tenth Machine Translation Summit', AAMT, AAMT, Phuket, Thailand, pp. 79–86.

Koehn, P. (2010), *Statistical Machine Translation*, Cambridge University Press.

Krynicki, G. (2006), Compilation, Annotation and Alignment of a Polish-English Parallel Corpus, PhD thesis, Poznan University.

Lukács, A., Pléh, C. & Racsmány, M. (2007), Spatial language in Williams syndrome: evidence for a special interaction?', *Journal of Child Language* **34(2):311-43.**

Magyar, L. & Szentgyörgyi, S. (2011), Vowel zero alternations in Hungarian nominal inflectional and derivational paradigms: An analogy-based statistical approach, *in* '4th Syntax, Phonology and Language Analysis Conference, Budapest'.

Manning, C. & Schütze, H. (1999), *Foundations of Statistical Natural Language Processing*, The MIT Press, Cambridge, MA.

Melamed, I. D. (1998), Empirical methods for exploiting parallel texts, PhD thesis, University of Pennsylvania, Philadelphia, PA, USA. AAI9829948.

Mihăltz, M. (2011), Magyar NP-felismerők összehasonlítása, *in* 'VIII. Magyar Számítógépes Nyelvészeti Konferencia. Szeged', pp. 333–335.

Mihăltz, M. & Pohl, G. (2006), Exploiting Parallel Corpora for Supervised Word-Sense Disambiguation in English-Hungarian Machine Translation, *in* 'Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006)', Genoa, Italy.

Mitkov, R. (2005), *The Oxford Handbook Of Computational Linguistics*, Oxford handbook, Oxford University Press.

Moore, R. C. (2002), Fast and accurate sentence alignment of bilingual corpora, *in* 'Proc 5th AMTA Conf: Machine Translation: From Research to Real Users', Springer, Langhorne, PA, pp. 135–244.

Oravecz, Cs. & Dienes, P. (2002), Efficient stochastic part-of-speech tagging for Hungarian, *in* 'Proceedings of the Third International Conference on Language Resources and Evaluation (LREC2002)', pp. 710–717.

Potthast, M., Barrón-Cedeño, A., Stein, B. & Rosso, P. (2011), Cross-language plagiarism detection', *Lang. Resour. Eval.* **45**(1), 45–62.

Prószéky, G., Tihanyi, L. & Ugray, G. (2004), Moose: a robust high-performance parser and generator, *in* 'Proceedings of the 9th Workshop of the European Association for Machine Translation', La Valletta, Malta, p. 138–142.

Racsmány, M., Conway, M., Keresztes, A. & Krajcsi, A. (2012), 'Inhibition and interference in the think/no-think task', *Memory and Cognition* **40(2):168-76.**

Rácz, P. & Szeredi, D. (2009), Testing usage-based predictions on Hungarian vowel reduction, *in* '17th Manchester Phonology Meeting, Manchester, UK'.

Recski, G., Rung, A., Zséder, A. & Kornai, A. (2010), NP Alignment in Bilingual Corpora, *in* N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner & D. Tapias, eds, 'Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)', European Language Resources Association (ELRA), Valletta, Malta.

Rosen, A. & Vavřín, M. (2012), Building a multilingual parallel corpus for human users, *in* N. C. C. Chair), K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, J. Odijk & S. Piperidis, eds, 'Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)', European Language Resources Association (ELRA), Istanbul, Turkey.

Solymosi, A. (2007), Tulajdonnév-felismerés, személynevek azonosítása magyar nyelvű szövegben, Master's thesis, Budapest University of Technology and Economics.

Szeredi, D. (2009), Functional phonological analysis of the Hungarian vowel system, Master's thesis, Eötvös Loránd University, Theoretical Linguistics.

Talvensaari, T. (2008), *Comparable Corpora in Cross-language Information Retrieval*, Julkaisusarja A, University of Tampere, Department of Computer Sciences.

Tiedemann, J. (2002), Uplug - a modular corpus tool for parallel corpora, *in* L. Borin, ed., 'Parallel Corpora, Parallel Worlds', Rodopi, Amsterdam, New York, pp. 181–197. Proceedings of the Symposium on Parallel Corpora, Department of Linguistics, Uppsala University, Sweden, 1999.

Tiedemann, J. (2009), News from OPUS – A Collection of Multilingual Parallel Corpora with Tools and Interfaces, *in* N. Nicolov, G. Angelova & R. Mitkov, eds, 'Recent Advances in Natural Language Processing V', Vol. 309 of *Current Issues in Linguistic Theory*, John Benjamins, Amsterdam & Philadelphia, pp. 227–248.

Turchi, M., Flaounas, I., Ali, O., Bie, T., Snowsill, T. & Cristianini, N. (2009), Found in Translation, *in* 'Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases: Part II', ECML PKDD '09, Springer-Verlag, Berlin, Heidelberg, pp. 746–749.

Váradi, T. (2002), The Hungarian National Corpus, *in* 'Proceedings of the Third International Conference on Language Resources and Evaluation', Las Palmas, pp. 385–389.

Váradi, T. (2003), Shallow parsing of hungarian business news, *in* 'Proceedings of Workshop on Shallow Processing of Large Corpora, March 27 (SProLaC03)', Lancaster, UK.

Váradi, T. & Gábor, K. (2004), A magyar intex fejlesztéséről, *in* Z. Alexin & D. Csendes, eds, 'II. Magyar Számítógépes Nyelvészeti Konferencia', Szegedi Tudományegyetem, Szeged, pp. 3–10.

Waldenfels, R. v. (2011), Recent Developments in Parasol: Breadth for Depth and Xslt Based Web Concordancing with Cwb, *in* 'Proceedings of Slovko 2011, Modra, Slovakia, 20–21 October 2011', pp. 156–162.